Statistical Guarantees for the EM Algorithm: From Population to Sample-Based Analysis

Authors: Balakrishnan, Wainwright, and Yu
Chenxi Zhou

Reading Group in Statistical Learning and Data Mining
September 5th, 2017

Outline

1️⃣ Overview of Expectation-Maximization (EM) Algorithm
2️⃣ Population Analysis of First-Order EM Algorithm
3️⃣ Sample Analysis of First-Order EM Algorithm
4️⃣ Example: Gaussian Mixture Model

Overview of Expectation-Maximization Algorithm

Estimation of Linkage in Genetics

- 197 animals are distributed multinomially into 5 categories
- Observed data:

$$\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) = (x_1, x_2, 18, 20, 34)$$

with $x_1 + x_2 = 125$.

- Cell probabilities:

$$\mathbf{p} = \left(\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi\right) \qquad \text{for } \pi \in [0,1].$$

Example (continued)

» Likelihood function:

$$L(\mathbf{p} \mid \mathbf{x}) = \frac{n!}{x_1!x_2!x_3!x_4!x_5!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{\pi}{4}\right)^{x_2}\left(\frac{1-\pi}{4}\right)^{x_3}\left(\frac{1-\pi}{4}\right)^{x_4}\left(\frac{\pi}{4}\right)^{x_5}$$

and

$$\hat{\pi}_{\mathsf{MLE}} = \frac{x_2 + x_5}{x_2 + x_3 + x_4 + x_5} = \frac{x_2 + 34}{x_2 + 18 + 20 + 34}$$

» How to solve this type of incomplete data problem?

What is the EM Algorithm?

Expectation-Maximization (EM) Algorithm is an *iterative* method that attempts to find the *maximum likelihood estimator* of a parameter θ of a *parametric* probability distribution in *incomplete data* problems.

Incompleteness:

- Missing data
- Censored or grouped data
- Latent class and latent data structures
- $\cdots$

Basic Setup

- Let $(Y, Z) \in \mathcal{Y} \times \mathcal{Z}$ with the joint density function f_{θ^*}, where
  - $f_{\theta^*} \in \{f_\theta : \theta \in \Omega\}$
  - Ω is non-empty, compact convex set
- Observe n i.i.d. copies of Y, $\{Y_1, \ldots, Y_n\}$
  - $Z_1, \ldots, Z_n$ are *missing* or *latent*
- **Goal:** Estimate θ^* by maximizing log-likelihood:

$$\ell_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \log\left(\int_{\mathcal{Z}} f_\theta(Y_i, z)\, dz \right)$$

EM Idea

Unfortunately, maximizing ℓ_n directly can be **hard**! But often the *complete data* log-likelihood

$$\frac{1}{n}\sum_{i=1}^{n}\log f_\theta(Y_i, Z_i)$$

is easier to maximize. So we replace the complete data log-likelihood by its conditional expectation:

$$Q_n(\theta \mid \theta^t) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}_{\theta^t}\left\{\log f_\theta(Y_i, Z_i) \mid Y_i\right\}$$

where expectation is computed with respect to current iterate θ^t.

EM Algorithm

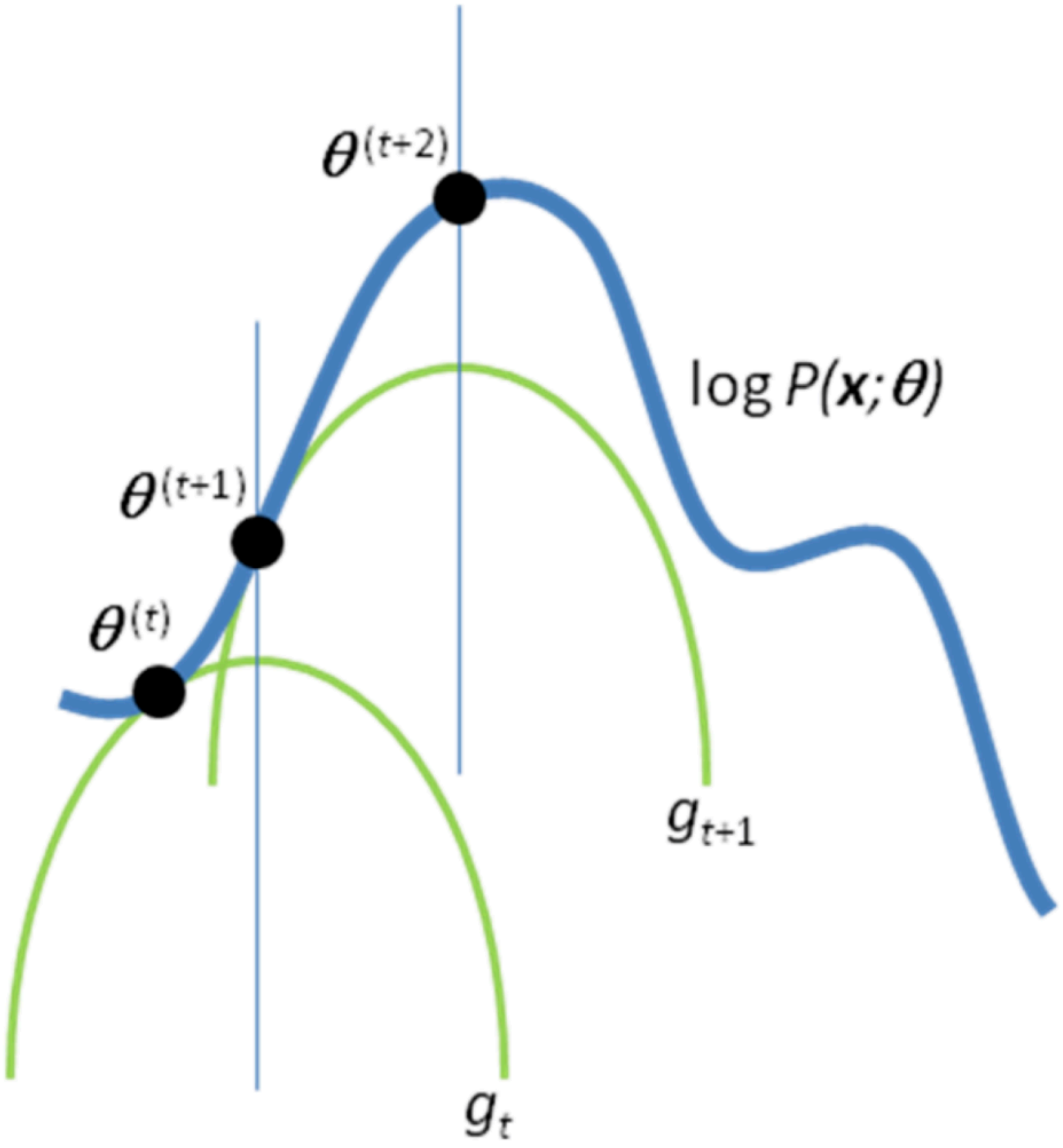
Starting with initial iterate $\theta^0 \in \Omega$, iterate the following steps for $t = 1, 2, \cdots$.

» **Expectation Step:** Compute EM surrogate $Q_n : \Omega \times \Omega \to \mathbb{R}$:

$$Q_n(\theta \mid \theta^t) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}_{\theta^t} \{\log f_\theta(Y_i, Z_i) \mid Y_i\}$$

» **Maximization Step:** Maximize EM surrogate:

$$\theta^{t+1} = \arg\max_{\theta \in \Omega} Q_n(\theta \mid \theta^t)$$



Source: `people.duke.edu/~ccc14/sta-663/EMAlgorithm.html`

Advantages 😃

- Easy to implement
- Requires small storage space
- Low cost per iteration
- If $\ell(\theta)$ is bounded, $\ell\left(\theta^t\right)$ converges monotonically to $\ell\left(\tilde{\theta}\right)$, where $\tilde{\theta}$ is a *stationary* point
- $\cdots$

Drawbacks 😔

» Finding the *exact* maximizer in the M step can be hard

As for the First Drawback...

» **Generalized EM Algorithm:** Just choose $\theta^{t+1} \in \Omega$ so that

$$Q_n\left(\theta^{t+1} \mid \theta^t\right) \geq Q_n\left(\theta^t \mid \theta^t\right).$$

» **First-Order EM Algorithm:** Assume $Q_n\left(\theta \mid \theta^t\right)$ is differentiable in the first argument at each iteration t. Given a step size $\alpha > 0$, the updates are

$$\theta^{t+1} = \theta^t + \alpha \cdot \nabla Q_n\left(\theta \mid \theta^t\right)\Big|_{\theta=\theta^t} \quad \text{for } t = 0, 1, 2, \cdots,$$

where the gradient is taken in the first argument of Q_n.

Drawbacks 😔

- Finding the *exact* maximizer in the M step can be hard
- No guarantees to converge to the global maximum of ℓ_n (depending on the choice of *starting point*)

An Example (Murray, 1977; Wu, 1983)

Twelve observations are collected from a bivariate normal distribution with mean $\mathbf{0}$, correlation coefficient ρ and variances σ_1^2, σ_2^2

Variable 1	1	1	−1	−1	2	2	−2	−2	*	*	*	*
Variable 2	1	−1	1	−1	*	*	*	*	2	2	−2	−2

The likelihood function has

- two global maxima: $\rho = \pm\frac{1}{2}, \sigma_1^2 = \sigma_2^2 = \frac{8}{3}$; and
- a saddle point: $\rho = 0, \sigma_1^2 = \sigma_2^2 = \frac{5}{2}$.

The EM algorithm starting at $\rho = 0$ will return the saddle point.

Drawbacks 😔

- Finding the *exact* maximizer in the M step can be hard
- No guarantees to converge to the global maximum of ℓ_n (depending on the choice of *starting point*)
- $\ell\left(\theta^t\right) \to \ell\left(\tilde{\theta}\right)$, where $\tilde{\theta}$ is a *stationary* point, does NOT imply $\theta^t \to \theta^*$ and Wu (1983) only established the conditions of convergence of $\left\{\theta^t\right\}_{t=1}^{\infty}$ to a *stationary* point

Contributions of Balakrishnan et al. (2017)

» Quantitative characterization of a basin of attraction around θ^*

» Where to choose the initialization to ensure $\theta^t \to \theta^*$

» Establishment of the convergence rate and the corresponding conditions

» Establishment of connections between population and sample analysis

Population Version of EM Algorithm

E Step: Compute the following population version surrogate function

$$Q\left(\theta \mid \theta^t\right) = \mathsf{E}_{\theta^*}\left\{\mathsf{E}_{\theta^t}\left(\log f_\theta(Y, Z) \mid Y\right)\right\}$$

M Step:

» *Standard EM:*

$$\theta^{t+1} = \underset{\theta \in \Omega}{\arg\max}\, Q\left(\theta \mid \theta^t\right)$$

» *First-Order EM:*

$$\theta^{t+1} = \theta^t + \alpha \cdot \nabla Q\left(\theta \mid \theta^t\right)\Big|_{\theta=\theta^t}$$

Oracle Surrogate Function and Iterates

» Oracle Surrogate Function

$$q(\theta) := Q(\theta \mid \theta^*) = \mathsf{E}_{\theta^*} \left\{ \mathsf{E}_{\theta^*} \left(\log f_\theta \left(Y, Z \right) \mid Y \right) \right\}$$

» Oracle Iterates

$$\tilde{\theta}^{t+1} = \tilde{\theta}^{t} + \alpha \cdot \nabla q \left(\tilde{\theta}^{t} \right)$$

Oracle Surrogate Function and Iterates

Why do we need them?

- If $q(\theta)$ satisfies *strong concavity* and *smoothness*, then the gradient ascent updates achieve *geometric convergence rate* to θ^*
- The population version first-order EM updates can be viewed as a perturbation of the oracle updates
- Therefore, in the population level analysis of EM algorithm, we need to control the quantity

$$\nabla q(\theta) - \nabla Q(\theta \mid \theta)$$

Population Analysis of First-Order EM Algorithm

Recall the population first-order updates are

$$\theta^{t+1} = \theta^t + \alpha \cdot \nabla Q \left(\theta \mid \theta^t\right)\Big|_{\theta=\theta^t},$$

and the oracle updates are

$$\theta^{t+1} = \theta^t + \alpha \cdot \nabla q \left(\theta^t\right).$$

Condition 1: Gradient Smoothness

For an appropriately *small* parameter $\gamma \geq 0$,

$$\|\nabla q(\theta) - \nabla Q\left(\theta \mid \theta\right)\|_2 \leq \gamma \|\theta - \theta^*\|_2$$

for all $\theta \in \mathbb{B}\left(r; \theta^*\right)$.

Condition 2: λ-Strong Concavity

There is some $\lambda > 0$ such that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \leq -\frac{\lambda}{2} \|\theta_1 - \theta_2\|_2^2$$

or, equivalently,

$$\langle \nabla q(\theta_1) - \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \leq -\lambda \|\theta_1 - \theta_2\|_2^2$$

for all pairs $\theta_1, \theta_2 \in \mathbb{B}(r; \theta^*)$.

Condition 3: μ-Smoothness

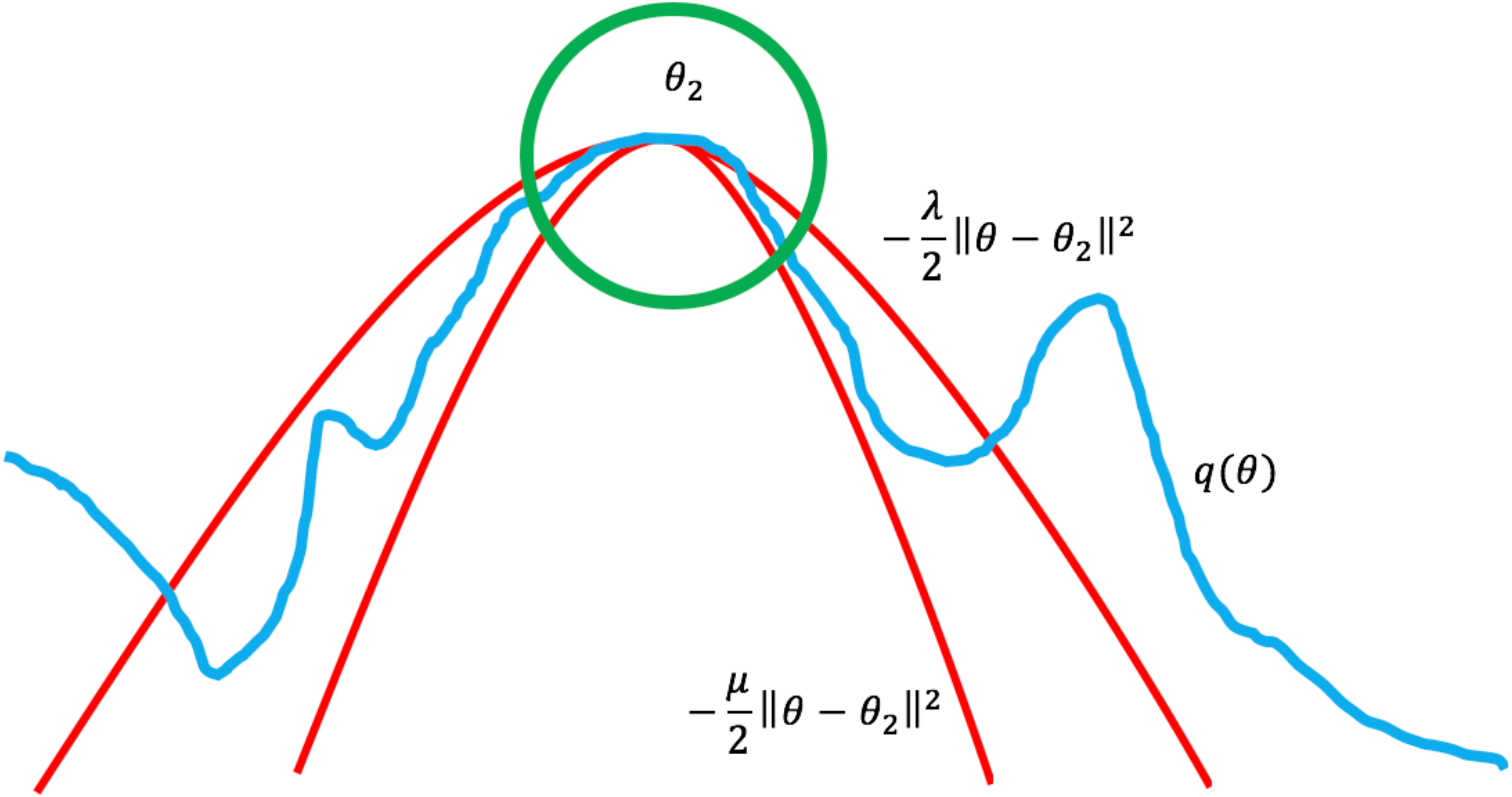
There is some $\mu > 0$ such that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \geq -\frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2$$

or, equivalently,

$$\langle \nabla q(\theta_1) - \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \geq -\mu \|\theta_1 - \theta_2\|_2^2$$

for all pairs $\theta_1, \theta_2 \in \mathbb{B}(r; \theta^*)$.


θ_2
$-\frac{\lambda}{2}\|\theta - \theta_2\|^2$
$q(\theta)$
$-\frac{\mu}{2}\|\theta - \theta_2\|^2$

Theorem 1

(General Population-Level Guarantee)

» For some radius $r > 0$ and a triplet (γ, λ, μ) with $0 \leq \gamma < \lambda \leq \mu$ such that γ-gradient smoothness, λ-strong concavity and μ-smoothness conditions hold;

» Choose the step size $\alpha = \dfrac{2}{\mu + \lambda}$.

Then, given any $\theta^0 \in \mathbb{B}(r; \theta^*)$, the population first-order EM iterates satisfy the bound

$$\left\|\theta^t - \theta^*\right\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)^t \left\|\theta^0 - \theta^*\right\|_2 \qquad \text{for all } t = 1, 2, \cdots.$$

Sample Analysis of First-Order EM Algorithm

Recall that the sample first-order EM updates are

$$\theta^{t+1} = \theta^t + \alpha \cdot \nabla Q_n \left(\theta \mid \theta^t\right)\Big|_{\theta=\theta^t}.$$

The analysis of the finite sample first-order EM algorithm depends on the empirical process

$$\left\{\nabla Q_n \left(\theta \mid \theta\right) - \nabla Q \left(\theta \mid \theta\right), \theta \in \mathbb{B}\left(r; \theta^*\right)\right\}.$$

For a given sample size n and tolerance parameter $\delta \in (0, 1)$, let $\epsilon_Q^{\text{unif}}(n, \delta)$ be the smallest scalar such that

$$\sup_{\theta \in \mathbb{B}(r; \theta^*)} \left\| \nabla Q_n(\theta \mid \theta) - \nabla Q(\theta \mid \theta) \right\|_2 \leq \epsilon_Q^{\text{unif}}(n, \delta)$$

with probability at least $1 - \delta$.

Theorem 2

(General Sample-Level Guarantee)

- For some radius $r > 0$ and a triplet (γ, λ, μ) with $0 \leq \gamma < \lambda \leq \mu$ such that the γ-gradient smoothness, λ-strong concavity and μ-smoothness conditions hold;
- Choose the step size $\alpha = \dfrac{2}{\mu + \lambda}$;
- Suppose the sample size n is large enough to ensure

$$\epsilon_Q^{\mathrm{unif}}(n, \delta) \leq (\lambda - \gamma) \cdot r.$$

Theorem 2

(General Sample-Level Guarantee) (continued)

Then, with probability at least $1-\delta$, given any initialization $\theta^0 \in \mathbb{B}(r;\theta^*)$, the finite-sample first-order EM iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy the bound

$$\left\|\theta^t-\theta^*\right\|_2 \leq \left(1-\frac{2\lambda-2\gamma}{\mu+\lambda}\right)^t\left\|\theta^0-\theta^*\right\|_2+\frac{\epsilon_Q^{\mathrm{unif}}(n,\delta)}{\lambda-\gamma}$$

for all $t=1,2,\cdots$.

Example: Gaussian Mixture Model

Consider the following two-component Gaussian mixture model

$$Y = \psi \cdot \theta^* + \epsilon,$$

where

$$\psi = \begin{cases} +1, & \text{w.p } \frac{1}{2} \\ -1, & \text{w.p } \frac{1}{2} \end{cases}, \qquad \epsilon \sim \mathcal{N}_d\left(0, \sigma^2 I\right),$$

and ψ and ϵ are independent.

Key Quantity:

$$\mathsf{SNR} = \frac{\|\theta^*\|_2}{\sigma}.$$

To analyze the EM algorithm at the *population* level, i.e., apply **Theorem 1**, one needs to establish the gradient smoothness, λ-strong concavity and μ-smoothness.

Oracle Function:

$$q(\theta) = -\frac{1}{2} \cdot \mathsf{E}_{\theta^*} \left\{ (1 - w_{\theta^*}(Y)) \cdot \|Y + \theta\|_2^2 + w_{\theta^*}(Y) \cdot \|Y - \theta\|_2^2 \right\},$$

where the weighting function $w_{\theta^*}(y)$ is a smooth function.

It is easy to verify that q is *strongly-concave* and *smooth* with parameters 1, i.e., $\lambda = \mu = 1$.

What about *gradient smoothness*?

Lemma 2

- Let $\mathsf{SNR} = \frac{\|\theta^*\|_2}{\sigma} \geq \eta$ for a sufficiently large $\eta > 0$;
- Let the radius be $r = \frac{\|\theta^*\|_2}{4}$.

Then, there is a constant $\gamma \in (0, 1)$ with $\gamma \leq \exp\left(-c_2 \eta^2\right)$ such that

$$\|\mathsf{E}\left\{2 \cdot \left(w_\theta(Y) - w_{\theta^*}(Y)\right) \cdot Y\right\}\|_2 \leq \gamma \cdot \|\theta - \theta^*\|_2 .$$

Corollary 1

(Population result for the first-order EM algorithm for GMM)

- Let $\mathsf{SNR} = \frac{\|\theta^*\|_2}{\sigma} \geq \eta$ for a sufficiently large $\eta > 0$;
- Let the radius $r = \frac{\|\theta^*\|_2}{4}$;
- Choose the step size $\alpha = 1$.

Corollary 1

(Population result for the first-order EM algorithm for GMM) (continued)

Then, there is a contraction coefficient $\kappa(\eta) \leq \exp\left(-c\eta^2\right)$, where c is a universal constant, such that for any initialization $\theta^0 \in \mathbb{B}\left(\frac{\|\theta^*\|_2}{4}; \theta^*\right)$, the population first-order EM iterates satisfy the bound

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2$$

for all $t = 1, 2, \cdots$.

Now, we go from the population to the sample-based analysis of this particular model.

At the *sample* level, we study the random variable

$$\|\alpha \cdot \nabla Q_n(\theta \mid \theta) - \alpha \cdot \nabla Q(\theta \mid \theta)\|_2$$
$$= \left\| \frac{1}{n} \sum_{i=1}^{n} (2w_\theta(Y_i) - 1) Y_i - \mathsf{E}(2 \cdot w_\theta(Y) Y - Y) \right\|_2$$

over the ball $\mathbb{B}\left(\frac{\|\theta^*\|_2}{4}; \theta^*\right)$.

Corollary 4

(Sample-based result for first-order EM guarantees for GMM)

- Let $\mathsf{SNR} = \dfrac{\|\theta^*\|_2}{\sigma} \geq \eta$ for a sufficiently large $\eta > 0$;
- Choose the radius $r = \dfrac{\|\theta^*\|_2}{4}$;
- Choose the step size $\alpha = 1$;
- Suppose the sample size n is lower bounded by $n \geq c_1 d \log(1/\delta)$.

Corollary 4

(Sample-based result for first-order EM guarantees for GMM)(continued)

Then, there is a contraction coefficient $\kappa(\eta) \leq \exp\left(-c\eta^2\right)$, where c is a universal constant, such that, for any initialization $\theta^0 \in \mathbb{B}\left(\frac{\|\theta^*\|_2}{4}; \theta^*\right)$, the first-order EM iterates $\{\theta_t\}_{t=0}^{\infty}$ satisfy the bound

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 + \frac{c_2}{1-\kappa} \|\theta^*\|_2 \left(1 + \frac{\|\theta^*\|_2^2}{\sigma^2}\right) \sqrt{\frac{d}{n} \log\left(\frac{1}{\delta}\right)},$$

with probability at least $1 - \delta$.

Summary

- This paper advances our *theoretical* understanding of EM algorithm.
- This paper concentrates on how to obtain a *near-optimal* estimate of θ^* using EM algorithm.
- With the help of *optimization theory*, this paper establishes the *size* of the region of attraction where the initialization should be chosen and the *rate of convergence* of the EM algorithm.
- This paper also develops techniques to analyze other algorithms for solving non-convex problems.
- What's next...


thank you
danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
mersi
barka
welalin
tack
spas
kia ora
vinaka
blagodaram
dank je
misaotra
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
nanni
nandri
dhanyavad
matondo
gracias
хвала
asante
mahana
obrigada
tenki
bedankt
bayarlalaa
gracie
hvala
mauruuru
koszonom
aciu
dankon
akun
djiere dieuf
lau
mochchakkeram
murakoze
enkosi
dziękuję
chnorakaloutioun
grafias ago
gracies
sulpay
mamnun
дякую
go raibh maith agat
chokrane
sobodi
dekuji
obrigado
sukriya
kop khun krap
taiku
arigatô
takk
dakujem
trugarez
mesi
didi madloba
sagolun
najis tuke
ありがとう
tanemirt
rahmet
grazie
terima kasih
kam sah hamnida
rahmat
diolch
dhanyavadagalu
shukriya
merce
мерси
감사합니다
xiexie
ευχαριστώ
merci